# Calidad y Preprocesamiento de Datos

MCIC VÍCTOR MANUEL CORZA VARGAS
MC FERNANDO AVITÚA VARELA

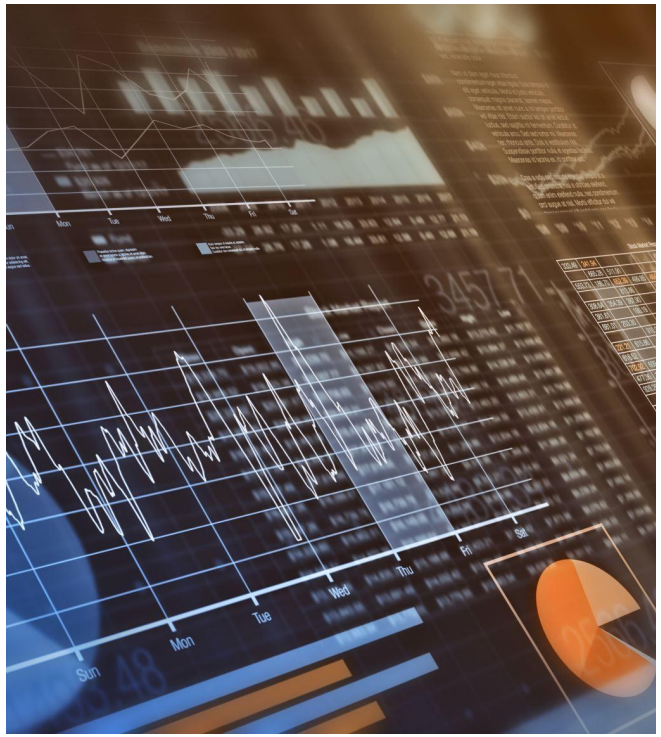
# Introducción a la Calidad de datos

# Introducción

**Objetivos:**

- Principales causas de problemas de calidad en bases de datos.
- Dimensiones de calidad de información y datos.
- Interdependencia de las dimensiones de calidad.

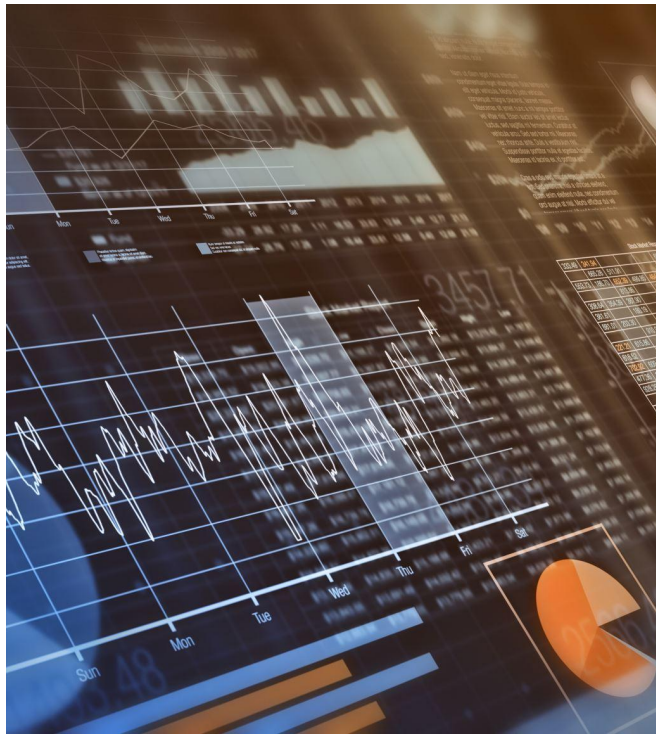
# Causas de Problemas de Calidad en las Bases de Datos




Los problemas de calidad en bases de datos pueden originarse por diversas razones, entre ellas:

- Errores Humanos: Ingresos incorrectos, duplicados, omisiones.
- Sistemas Heterogéneos: Integración de datos de diferentes fuentes sin estandarización.
- Problemas de Migración: Conversión incorrecta de datos entre sistemas.

# Causas de Problemas de Calidad en las Bases de Datos




- Datos Obsoletos: Falta de actualización de la información.
- Falta de Estándares: Ausencia de reglas claras para la gestión de datos.
- Errores de Captura Automática: Fallas en sensores, OCR, procesos ETL defectuosos.

# Calidad de Información

Se refiere a la utilidad de los datos en un contexto específico. Los principales factores que influyen en la calidad de la información incluyen:

- Relevancia: La información obtenida a partir de los datos disponibles debe ser útil y aplicable a la situación.
- Oportunidad: La información debe estar disponible en el momento adecuado.
- Exactitud: La información debe representar la realidad con precisión.
- Interpretabilidad: La información debe ser comprensible y utilizable por los usuarios.

# Calidad de Datos

En el ámbito de la calidad de datos y de la información se ha adoptado la palabra dimensión para identificar aquellos aspectos de los datos que pueden ser medidos y a través de los cuales su calidad puede ser cuantificada. Entre las principales dimensiones de calidad de datos encontramos:

**Exactitud**: Los datos reflejan la realidad sin errores.

**Completitud**: No hay datos faltantes o vacíos.

**Consistencia**: Los datos mantienen coherencia a lo largo del sistema.

**Validez**: Cumplen con las *reglas* y formatos esperados.

**Accesibilidad**: Disponibilidad fácil para los usuarios autorizados.

**Trazabilidad**: Se puede rastrear el origen y transformaciones de los datos.

# Interdependencia de Dimensiones de Calidad de Datos

Las dimensiones de calidad de datos no son independientes, sino que interactúan entre sí de diversas maneras:

**Relación entre Exactitud y Consistencia**: Datos exactos pero inconsistentes pueden generar confusión en los análisis.

# Interdependencia de Dimensiones de Calidad de Datos

**Relación entre Exactitud y Consistencia**: Datos exactos pero inconsistentes pueden generar confusión en los análisis.

**Puedes tener dos bases de datos con información real (exacta), pero si una usa "EE.UU." y la otra "Estados Unidos", tus reportes se duplicarán.**

# Interdependencia de Dimensiones de Calidad de Datos

**Completitud y Validez**: Un dato puede estar presente (completo) pero no cumplir con los requisitos esperados (no válido).

# Interdependencia de Dimensiones de Calidad de Datos

**Completitud y Validez**: Un dato puede estar presente (completo) pero no cumplir con los requisitos esperados (no válido).

**Este es el clásico caso de "rellenar por rellenar". Un campo de "Teléfono" puede estar lleno (completo), pero si tiene el texto "12345", no es válido.**

**La completitud sin validez es solo ruido. Los datos inválidos son más peligrosos que los datos faltantes porque pueden sesgar los algoritmos sin que te des cuenta.**

# Interdependencia de Dimensiones de Calidad de Datos

**Oportunidad vs. Exactitud:** A veces, para que un dato sea 100% exacto, tardamos demasiado en procesarlo.

En el mundo real, a veces preferimos un dato 95% exacto *ahora*, que uno 100% exacto dentro de un mes.

# Interdependencia de Dimensiones de Calidad de Datos

**Accesibilidad y Seguridad**: Si bien los datos deben ser accesibles, también deben cumplir con requisitos de seguridad para evitar el acceso no autorizado.

Este es el "Gran Dilema" del Gobierno de Datos.

Refinamiento: Más que una relación, es una tensión equilibrada. Demasiada seguridad mata la agilidad; demasiada accesibilidad compromete la privacidad. (Para equilibrar se usan frameworks como GDPR)

# General Data Protection Regulation

Es la ley de **privacidad y seguridad** más estricta del mundo. Fue creada por la Unión Europea (entró en vigor en 2018), pero tiene un alcance **global**.

**¿Por qué afecta a las empresas aunque no estén en Europa?**

Aquí está el truco: si una empresa (aunque esté en México, Argentina o EE. UU.) ofrece productos o servicios a personas en la Unión Europea, o simplemente recolecta sus datos, **debe cumplir con el GDPR**. Por eso ves avisos de "Aceptar Cookies" en casi todas las webs del mundo.

# Puntos clave del GDPR

El GDPR es relevante porque impone reglas sobre cómo se manejan las dimensiones de calidad de datos:

- **Consentimiento:** Los usuarios deben dar permiso explícito y claro para que se usen sus datos. No son válidas las casillas pre-marcadas.
- **Derecho al olvido:** Si un usuario lo pide, la empresa debe borrar todos sus datos personales.
- **Portabilidad:** El usuario tiene derecho a pedir que le entreguen sus datos en un formato fácil de leer para llevarlos a otro lado.
- **Multas masivas:** El incumplimiento puede costar hasta **20 millones de euros** o el **4% de la facturación global** anual de la empresa (lo que sea más alto).

# Puntos clave del GDPR

Gracias al GDPR, muchas empresas tuvieron que limpiar sus bases de datos, mejorando drásticamente su **Calidad de Datos**, ya que la ley exige que los datos sean "exactos y, si es necesario, actualizados".

# 2.3 Modelo de Medición de Calidad de Datos

Para medir la calidad de datos, se emplean métricas y técnicas específicas:

Métricas básicas: **Precisión** (% de valores correctos), **completitud** (% de datos sin valores nulos), **consistencia** (comparación entre fuentes), **oportunidad** (Tiempo necesario para que los datos estén disponibles)

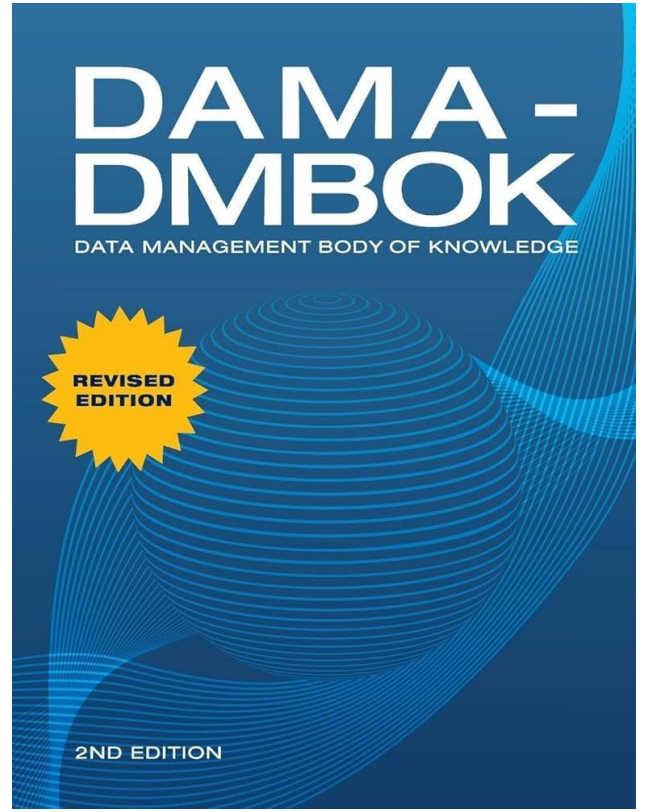
Métodos de evaluación: Encuestas, auditorías de datos, validaciones automáticas.

Herramientas de calidad de datos: Talend Data Quality, Informatica Data Quality, varias librerías de python: missingno, ydata_profiling, pandas, seaborn.

# 2.4 Modelos de Evaluación de Calidad de Datos

# Modelo de Gestión de Datos DAMA

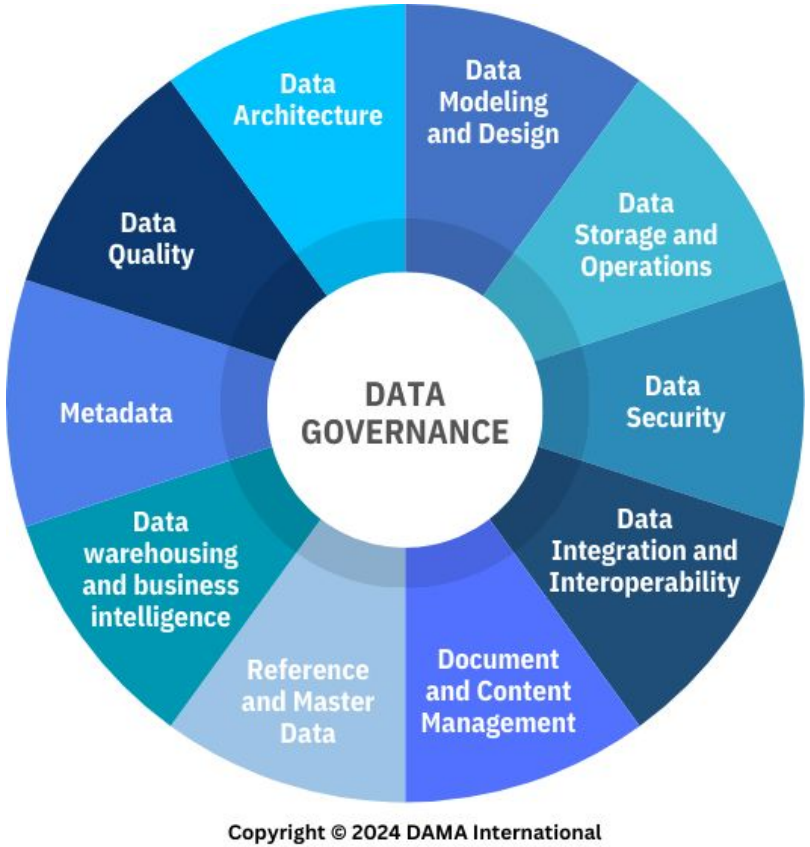
Un estándar de la gestión de datos empresariales que incluye estrategias para calidad de datos.



# Modelo de Gestión de Datos DAMA

11 áreas son las que componen la famosa **"Rueda de DAMA"** (DAMA Wheel). Es el mapa completo para cualquier organización que quiera pasar de "tener datos" a "gestionar activos de información".

Lo más importante es que el **Gobierno de Datos** está en el centro, porque es el eje que conecta y da dirección a todas las demás.




https://dama.org/dmbok2r-infographics/

# Las 11 Áreas de Conocimiento de DAMA-DMBOK2

**1. Gobierno de Datos (Data Governance):** El corazón del modelo. Se encarga de la estrategia, el control y la autoridad sobre la gestión de los datos.

**2. Arquitectura de Datos (Data Architecture):** Define el diseño maestro para gestionar los activos de datos y alinearlos con la estrategia de la empresa.

**3. Modelado y Diseño de Datos (Data Modeling & Design):** El proceso de descubrir, analizar y representar los requisitos de datos en modelos precisos.

**4. Almacenamiento y Operaciones de Datos (Data Storage & Operations):** La gestión del despliegue y soporte de los datos almacenados (DBA, mantenimiento de bases de datos).

# Las 11 Áreas de Conocimiento de DAMA-DMBOK2

**5. Seguridad de Datos (Data Security):** Asegurar la privacidad, confidencialidad y el acceso adecuado a los datos (aquí es donde entra el **GDPR** que mencionamos antes).

**6. Integración e Interoperabilidad de Datos (Data Integration & Interoperability):** Los procesos para mover y consolidar datos entre sistemas y aplicaciones.

**7. Gestión de Documentos y Contenido (Document & Content Management):** Gestión de datos no estructurados (PDFs, imágenes, correos) y su ciclo de vida.

**8. Datos Maestros y de Referencia (Reference & Master Data):** Gestión de los datos compartidos (como "Clientes" o "Productos") para reducir la redundancia y mejorar la consistencia.

# Las 11 Áreas de Conocimiento de DAMA-DMBOK2

**9. Data Warehousing y Business Intelligence (DW/BI):** La gestión del acceso a los datos para informes, análisis y toma de decisiones.

**10. Gestión de Metadatos (Metadata Management):** Gestionar la "información sobre la información" para que los datos sean fáciles de encontrar y entender.

**11. Gestión de la Calidad de Datos (Data Quality Management):** Asegura que los datos sean aptos para su uso.

# 2.4 Modelo de Evaluación de Calidad de Datos

Para evaluar la calidad de datos en un entorno organizacional, se siguen estos pasos:

1. **Definición de criterios de calidad**: Identificación de las dimensiones más relevantes para el contexto.
2. **Recolección de datos**: Uso de herramientas de análisis para obtener información sobre la calidad.
3. **Análisis de resultados**: Identificación de problemas y sus causas.
4. **Plan de mejora**: Estrategias para mejorar la calidad, como limpieza de datos, integración de fuentes o estandarización.
5. **Monitoreo continuo**: Implementación de controles y procesos automáticos para garantizar la mejora continua.

# Ejemplo de Uso del Modelo DAMA-DMBOK en Evaluación y Mejora de Calidad de Datos

# Contexto del Problema

Una institución financiera detecta problemas con la calidad de datos en su sistema de gestión de clientes. Se han identificado registros con información inconsistente en nombres, direcciones y datos de cuentas, lo que genera errores en la comunicación con clientes y en la asignación de créditos.

# 1 Medición de Calidad de Datos

Se utilizan métricas específicas para evaluar la calidad de los datos en la base de clientes.

| Dimensión de Calidad | Métrica | Resultado | Interpretación |
|---|---|---|---|
| Precisión | % de datos incorrectos en nombres o direcciones | 12% | 12% de los registros tienen errores ortográficos o de formato |
| Consistencia | % de cuentas con datos incoherentes en distintos sistemas | 18% | Datos no coinciden entre la base de clientes y el sistema de facturación |
| Completitud | % de registros con campos vacíos | 9% | 9% de los clientes no tienen número de teléfono registrado |
| Oportunidad | Tiempo promedio para actualizar datos incorrectos | 48 horas | Se tarda hasta 2 días en corregir información tras detectar un error |

# Evaluación y Diagnóstico

La empresa analiza los resultados de la medición y detecta los siguientes problemas:

- **Nombres y direcciones incorrectas** → Impactan la comunicación con clientes y procesos de verificación de identidad.
- **Datos inconsistentes entre sistemas** → Generan errores en la facturación y aprobaciones de crédito.
- **Falta de datos clave** → Dificulta la comunicación con clientes.
- **Tiempos de corrección lentos** → Afectan la eficiencia del servicio al cliente.

# Aplicación del Modelo DAMA-DMBOK

DAMA-DMBOK 2 define 11 áreas de conocimiento para la gestión de datos, para abordar la calidad de los datos utilizaremos cuatro de estas áreas:

1. Gestión de la Calidad de Datos (Data Quality Management)
2. Gestión de Datos Maestros (Master Data Management - MDM)
3. Gobierno de Datos (Data Governance)
4. Integración e Interoperabilidad de Datos

# 1. Gestión de la Calidad de Datos (Data Quality Management)

Este es el proceso central. DAMA lo define como el ciclo de planificar, controlar y mejorar la calidad de los datos.

- **Perfilado de Datos (Data Profiling):** Fue lo que se hizo para detectar ese **12% de errores en nombres** y el **9% de campos vacíos**. Consiste en inspeccionar los datos para entender sus fallas.
- **Limpieza y Mejora:** Aplicamos reglas de validación en los formularios para evitar que entre "basura" al sistema.

# Gestión de Datos Maestros (Master Data Management - MDM)

Este proceso es el que soluciona el **18% de inconsistencia** entre la base de clientes y el sistema de facturación.

- **Identificación del "Golden Record":** DAMA busca crear una versión única y confiable de un cliente. Si el sistema de facturación y el CRM (Customer Relationship Management) dicen cosas distintas, el MDM decide cuál es la "verdad".
- **Resolución de Entidades:** Es el proceso técnico para entender que "Juan Pérez" en el sistema A es el mismo "J. Pérez" en el sistema B.

# 3. Gobierno de Datos (Data Governance)

Este no es un proceso técnico, sino de **autoridad y rendición de cuentas**. En el caso de la institución bancaria se utiliza para atacar el retraso de **48 horas** en las correcciones.

- **Data Stewardship:** Se asignan responsables (Custodios de Datos). Sin un Steward, nadie se siente responsable de bajar esas 48 horas.
- **Establecimiento de SLAs (Acuerdos de Nivel de Servicio):** El Gobierno de Datos define que un error crítico *debe* corregirse en menos de X horas, convirtiendo la "buena voluntad" en una norma institucional.

# 4. Integración e Interoperabilidad de Datos

Este proceso se encarga del **flujo de los datos** entre sistemas.

- **Sincronización:** Para resolver la inconsistencia, DAMA sugiere pasar de procesos manuales o por lotes (que causan desfases) a integraciones en tiempo real o semi-real. Esto asegura que si cambias una dirección en un sistema, se refleje instantáneamente en el resto de la institución financiera.

# Otros marcos de trabajo para la evaluación de la Calidad de los Datos

Existen otros marcos de trabajo para evaluar la calidad de los datos, entre ellos encontramos:

- **Total Data Quality Management (TDQM):** A framework that manages data quality from collection to analysis. It was developed by Richard Y. Wang at MIT.
- **Data Quality Assessment Framework (DQAF):** A framework that defines standards, processes, and tools to ensure data quality. It also establishes guidelines for data governance. Utilizado en contextos gubernamentales para evaluar la calidad de datos estadísticos.
- **Data Quality Maturity Model (DQMM):** A framework that provides a roadmap for organizations to improve their data quality management practices.

# Otros marcos de trabajo para la evaluación de la Calidad de los Datos

- **Data Quality Scorecard (DQS):** A framework that uses statistical approaches to identify and eliminate data issues. It also provides a structured approach to evaluating and monitoring data quality.
- **Six Sigma:** A quality management methodology that uses statistical tools and techniques to identify and eliminate sources of data errors.
- **Root Cause Analysis:** A framework that identifies the underlying causes of a data-related problem.
- **Modelo de Wang y Strong**: Define dimensiones clave como precisión, coherencia y completitud.

# Recomendaciones para implementar una verificación de Calidad de datos

# Verificación de Calidad de Datos

**Para verificar la dimensión de Precisión**

- **Verificación manual**: Realiza una comparación de los datos con fuentes confiables o datos originales. Por ejemplo, si estás trabajando con datos de ventas, compáralos con los registros financieros oficiales.
- **Pruebas de consistencia**: Asegúrate de que no haya contradicciones en los datos. Por ejemplo, un mismo cliente no debe tener dos direcciones de envío contradictorias.
- **Algoritmos de validación**: Usa herramientas automáticas para detectar valores atípicos o errores numéricos.

# Verificación de Calidad de Datos

**Para verificar la dimensión de Completitud**

- **Revisión de valores faltantes**: Verifica que no haya campos vacíos en los registros que son clave para tu análisis. Existe una amplia diversidad de herramientas para detectar registros con valores nulos o vacíos (SQL, pandas(python), Informatica, Apache Hop, etc.)
- **Validación de registros**: Compara tus datos con un conjunto de datos de referencia para asegurarte de que los datos esperados están presentes.
- **Imputación de datos faltantes**: Si encuentras datos faltantes, puedes rellenarlos mediante técnicas de imputación, dependiendo del tipo de datos.

# Verificación de Calidad de Datos

**Para verificar la dimensión de Consistencia**

- **Reglas de validación**: Establece reglas que aseguren la coherencia. Por ejemplo, si una columna contiene códigos de país, todos los registros deben ajustarse a un conjunto de códigos válidos.
- **Análisis cruzado**: Verifica que los datos sean consistentes cuando se cruzan entre diferentes bases de datos o sistemas.
- **Detección de duplicados**: Usa herramientas para identificar registros duplicados o inconsistentes, como buscar nombres duplicados, fechas incorrectas o IDs repetidos.

# Verificación de Calidad de Datos

**Para verificar la dimensión de Oportunidad (en el aspecto de Actualización)**

- **Revisión periódica**: Establece procedimientos para revisar y actualizar los datos de manera regular, especialmente si estás utilizando datos en tiempo real o dinámicos (como precios de productos o inventarios).
- **Establecimiento de ciclos de actualización**: Si los datos se actualizan automáticamente, verifica la frecuencia con la que esto sucede y asegúrate de que los cambios se reflejan correctamente.
- **Auditorías de datos**: Realiza auditorías periódicas para confirmar que los datos se han actualizado correctamente en el sistema.

# Otras dimensiones

**Para verificar la dimensión de Fiabilidad**

- **Verificación de fuentes**: Asegúrate de que las fuentes de los datos sean confiables y estén bien establecidas. Por ejemplo, si obtienes datos de una API, verifica que esa API esté documentada y sea mantenida activamente.
- **Evaluación de origen de datos**: Identifica el origen de los datos para confirmar que provienen de fuentes legítimas y verificables.
- **Auditoría de la recopilación de datos**: Realiza auditorías para asegurar que los procesos de captura de datos sean correctos y que no haya errores introducidos durante la recopilación.

# Otras dimensiones

**Para verificar la dimensión de Relevancia**

- **Evaluación del contexto**: Pregúntate si los datos que estás utilizando son realmente relevantes para el propósito que deseas. Asegúrate de que los datos estén alineados con los objetivos de tu análisis.
- **Eliminación de datos innecesarios**: Si los datos no son útiles o no aportan valor, considera eliminarlos del conjunto para evitar que generen ruido en los análisis.
- **Conformidad con estándares del sector**: Si trabajas en un área específica (finanzas, salud, marketing, etc.), asegúrate de que los datos que manejas cumplan con los estándares de calidad del sector.

# Otras dimensiones

**Para verificar la dimensión de Accesibilidad (Podría ser otro aspecto de la dimensión de oportunidad)**

- **Revisión de permisos y acceso**: Verifica que los datos estén accesibles para las personas que los necesiten, sin barreras innecesarias. Asegúrate de que las políticas de privacidad o de seguridad no impidan el acceso a los datos.
- **Facilidad de acceso**: Evalúa si los datos están disponibles en formatos que sean fáciles de usar (como CSV, JSON, bases de datos SQL) y que puedan ser procesados sin problemas por las herramientas que uses.
- **Pruebas de accesibilidad**: Realiza pruebas para asegurarte de que los datos puedan ser consultados y utilizados sin demoras o dificultades técnicas.

# Algoritmos y herramientas en Calidad de datos

# Algoritmos y herramientas en calidad de datos

**Detección de Duplicados**

**Algoritmos de comparación exacta**: Estos algoritmos verifican la existencia de registros duplicados mediante una comparación exacta de todos los campos. Usualmente se utilizan en bases de datos para identificar registros duplicados.

- **Método**: Comparación de todos los valores de los registros (campos de texto, numéricos, etc.) para ver si son exactamente iguales.
- **Herramientas comunes**: SQL `DISTINCT`, Pandas (Python).

# Verificación de Calidad de Datos

**Algoritmos de comparación aproximada (Fuzzy Matching)**: Cuando los registros no son exactamente iguales, pero contienen errores tipográficos o variaciones, se puede usar un enfoque de coincidencia difusa (fuzzy matching).

- **Método**: Algoritmos como **Levenshtein Distance** (distancia de edición) o **Jaro-Winkler** miden la similitud entre cadenas de texto.
- **Herramientas comunes**: FuzzyWuzzy (Python), RecordLinkage (Python).

# Verificación de Calidad de Datos

**Detección de Valores Atípicos**

- **Algoritmos estadísticos**: Los valores atípicos (outliers) son aquellos que se desvían significativamente del comportamiento esperado de los datos. Algunos algoritmos que ayudan en esta tarea incluyen:
  - **Z-Score**: Identifica outliers basados en la desviación estándar de los datos. Si el valor de una observación tiene un Z-score superior a un umbral determinado (por ejemplo, ±3), se considera un outlier.
  - **IQR (Interquartile Range)**: Usa el rango intercuartílico para identificar valores atípicos. Los datos fuera del rango de Q1 - 1.5*IQR y Q3 + 1.5*IQR son considerados outliers.
  - **Método de Boxplot**: Utiliza un gráfico de caja (boxplot) para identificar los valores atípicos visualmente.

# Verificación de Calidad de Datos

**Detección de Valores Atípicos**

- **Modelos de aprendizaje automático**:
  - **Isolation Forest**: Un algoritmo de aprendizaje no supervisado que es eficiente para detectar valores atípicos en grandes conjuntos de datos.
  - **Local Outlier Factor (LOF)**: Detecta outliers al comparar la densidad local de los puntos con sus vecinos.

# Referencias


- Redman, T. C. (1996). Data Quality: The Field Guide. Digital Press.
- Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4), 5-33.
- Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211-218.
- Batini, C., & Scannapieco, M. (2016). Data Quality: Concepts, Methodologies and Techniques. Springer.
- Kimball, R., & Caserta, J. (2004). The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley.